Gene Mapping and Marker Clustering Using Shannon's Mutual Information

Zaher Dawy, Bernhard Goebel, Joachim Hagenauer,
Christophe Andreoli, Thomas Meitinger, and Jakob C. Mueller

Abstract—Finding the causal genetic regions underlying complex traits is one of the main aims in human genetics. In the context of complex diseases, which are believed to be controlled by multiple contributing loci of largely unknown effect and position, it is especially important to develop general yet sensitive methods for gene mapping. We discuss the use of Shannon's information theory for population-based gene mapping of discrete and quantitative traits and for marker clustering. Various measures of mutual information were employed in order to develop a comprehensive framework for gene mapping analyses. An algorithm aimed at finding so-called relevance chains of causal markers is proposed. Moreover, entropy measures are used in conjunction with multidimensional scaling to visualize clusters of genetic markers. The relevance chain algorithm successfully detected the two causal regions in a simulated scenario. The approach has also been applied to a published clinical study on autoimmune (Graves') disease. Results were consistent with those of standard statistical methods, but identified an additional locus of interest in the promotor region of the associated gene CTLA4. The developed software is freely available at http://www.lnt.ei.tum.de/download/InfoGeneMap/.

Index Terms—Complex traits, genotype-phenotype association, information theory, relevance chains, SNPs.

1 Introduction

One of the main aims of current genetics research is to discover functional connections between genotype and phenotype. Identifying the causal genetic variants and their functional patterns may greatly facilitate the (preventive) diagnosis and biochemical understanding of genetic diseases. This so-called *gene mapping* is especially important in the context of common complex disorders such as neurodegenerative or cardiovascular diseases. Complex diseases are believed to be caused by multiple contributing genetic loci with varying effect strengths and epistatic or interacting effects between markers. Moreover, the effect strength of a single causal factor can be quite small and the risk of developing the disease may also depend on environmental factors. These properties make the gene mapping of complex diseases a particularly difficult task [1].

With the development of rapid and cost-effective genotyping methods, the focus of research is shifting toward population-based case-control studies [2]. These studies usually investigate sequences of *single nucleotide polymorphisms (SNPs)*, which are the predominant form of polymorphisms in the human genome. Standard statistical methods have proven quite successful in identifying susceptibility regions for several traits ([1], [3], [4]) and citations therein). However, some of the traditional association testing methods, such as the allelic tests or trend tests, implicitly or explicitly make certain assumptions (e.g., Hardy-Weinberg equilibrium (HWE), multiplicative effects of alleles on the trait's penetrance, etc.), which may limit their usefulness to specific models of complex traits. To begin with a screen for potential genotype-phenotype associations in large data sets without assuming a specific risk model, it will be necessary to come up with general methods that systematically analyze the full genetic information provided by DNA variation with respect to the complex disease under investigation.

This notion of gene mapping as the task of analyzing genetic information has led to the idea of applying the methods of *information theory*. Information theory was established in 1948 by Shannon as a mathematical theory of communication [5]. While it is most commonly encountered in the context of communications engineering, information theory provides a general framework for the quantitative analysis of information, with applications in many fields of research, e.g., physics, statistics, economics, and engineering, to name but a few [6]. The basic concepts of information theory have been applied to various problems in molecular biology recently, mainly in the context of data mining. Examples include the use of mutual information to extract clusters of genes from RNA expression data [7] and of relative entropy (or Kullback-Leibler distance) to analyze patterns of gene expression [8]. In [9], a multilocus linkage disequilibrium measure is derived from Shannon's entropy; [10] uses entropy to select markers of interest for association studies. Grosse et al. propose a

- *Z. Dawy, B. Goebel, and J. Hagenauer are with the Institute for Communications Engineering (LNT), Munich University of Technology (TUM), Arcisstr. 21, 80290 Munich, Germany. Z. Dawy is also with the Department of Electrical and Computer Engineering, American University of Beirut, Riad El Solh, Beirut 1107 2020, Lebanon. E-mail: zaher.dawy@aub.edu.lb, {bernhard.goebel, hagenauer}@tum.de.*
- *C. Andreoli, T. Meitinger, and J.C. Mueller are with the Institute of Human Genetics, GSF-National Research Center for Environment and Health, 85764 Neuherberg, Germany. J.C. Mueller is also with the Institute for Medical Statistics and Epidemiology and the Institute for Psychiatry and Psychotherapy, Munich University of Technology (TUM), Ismaninger Str. 22, 81675 Munich, Germany. E-mail: christopheandreoli@yahoo.de, meitinger@gsf.de, jakob.mueller@imse.med.tu-muenchen.de.*

Manuscript received 2 Dec. 2004; revised 8 July 2005; accepted 28 July 2005; published online 31 Jan. 2006.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0198-1204.

successful application of mutual information to distinguish coding and noncoding DNA [11]. In [12], a very short introduction to the calculation of mutual information scores between SNPs and some partition (e.g., case-control groups) is presented.

In many examples found in the literature, information theoretic measures such as entropy, relative entropy, and mutual information are used to derive auxiliary measures seemingly appropriate to the problem at hand. The more advanced properties and theorems of information theory are, however, rarely exploited. As these properties are usually not inherited to derived methods and measures, the assessment of results often needs to be performed empirically.

In this paper, we apply information theoretic concepts to two classes of problems: 1) the gene mapping of complex diseases in population-based case-control studies and the mapping of quantitative traits and 2) the clustering of genetic markers in terms of their variability patterns. Both problems are interconnected in the context of complex traits because a trait-associated cluster of markers delineates the genomic region in which the true causal variant has to be searched for. Groups of genetic markers with similar variability patterns are suspected to have the same evolutionary history and will be jointly interpreted. In recognition of the complex genetic properties of the traits under investigation, the main premise of our work was to develop a general method for the analysis of directly measured marker genotypes. We successfully tested our method on both simulated data sets and clinical studies.

The paper is organized as follows: Section 2 briefly reviews some basic principles from information theory. Section 3 works out methods for one-locus and multiple-loci gene mapping using Shannon's mutual information. Moreover, it presents a relevance chains algorithm and an extension to data sets with continuous phenotypes. The use of information theory is extended to the problem of marker clustering in Section 4. The different proposed methods are then tested on simulated and clinical data sets and results are presented in Section 5. Finally, conclusions are drawn in Section 6.

2 Information Theory

This section gives some important definitions of basic concepts from information theory. A more exhaustive treatment of the topic can be found in [6].

At the heart of information theory is the concept of *entropy*. This term was coined following the notion of entropy in physics and characterizes the quantity of a random process's uncertainty. The entropy of a random variable X with realizations x is defined as (for notes on the notation used, see Appendix A)

$$H(X) = -\sum_x p(x)\log_2 p(x). \tag{1}$$

Most commonly, the base 2 logarithm is used, leading to a result in unit *bits*. However, other bases may be used as well, e.g., the natural logarithm, where the unit is *nats*. The conditional entropy of a random variable X given another random variable Y is

$$H(X|Y) = -\sum_x \sum_y p(x,y)\log_2 p(x|y). \tag{2}$$

In other words, conditional entropy $H(X|Y)$ is the entropy of X that remains when Y is observed. With these definitions, the concept of *mutual information* (MI), defined as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_x \sum_y p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}, \end{aligned} \tag{3}$$

can be understood as the reduction in entropy (or uncertainty) of one random variable given another. Mutual information has applications to many fields of science; it is, for instance, used to calculate the channel capacity in communications engineering. In this paper, it will be used as a measure of dependence between phenotype and genotype. This is motivated by the property

$$I(X;Y) = I(Y;X) \geq 0, \tag{4}$$

with equality if and only if X and Y are statistically independent. Other definitions required for the following sections are those of *joint* and *conditional mutual information*, given as

$$\begin{aligned} I(X,Y;Z) &= H(X,Y) - H(X,Y|Z) \\ &= \sum_x \sum_y \sum_z p(x,y,z)\log_2 \frac{p(x,y,z)}{p(x,y)p(z)}, \end{aligned} \tag{5}$$

and

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \sum_x \sum_y \sum_z p(x,y,z)\log_2 \frac{p(x,y|z)}{p(x|z)p(y|z)}. \end{aligned} \tag{6}$$

These are connected by the *chain rule of mutual information*,

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X). \tag{7}$$

For later sections, it is important to note that the random variables in (5) and (6) may be vector random variables combining two or more single variables.

3 One-Locus and Multiple-Loci Gene Mapping

Given a case-control study comprised of N individuals and an SNP sequence of length L, we conceive the phenotype as well as the individual SNPs as random variables P and $S_1, S_2, \ldots, S_L$, respectively. Usually, P will be a binary variable (case or control) and each S_i will be a ternary variable (two homozygous and one heterozygous combinations of two alleles in a diploid organism). The probabilities of the random variables' states can be derived from relative frequencies, i.e., observed counts divided by N. These probability estimates exhibit a variance that depends on the sample size N. This issue will be addressed in the context of significance evaluation.

To investigate each SNP's causality, we calculate the mutual information between the SNP and the phenotype, $I(S_i;P), i = 1 \ldots L$ [13]. In a case-control study with an equal number of cases and controls, $H(P) = 1\ \text{bit}$ and,

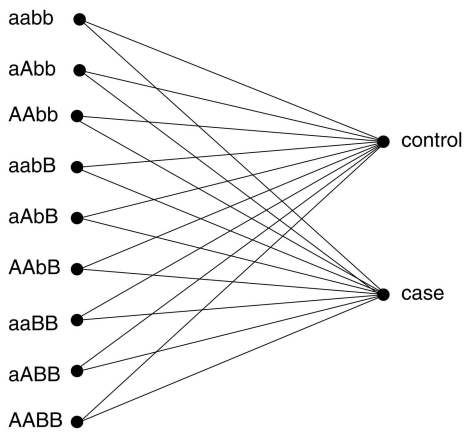

Fig. 1. Genotype-phenotype transition diagram for a two-locus model.

therefore, $0\text{ bit} \leq I(S;P) \leq 1\text{ bit}$. If a higher number of jointly analyzed markers is suspected, the model can be extended to this case by forming a vector random variable $\mathbf{S}$ combining a number of SNPs, $\mathbf{S} = (S_i, S_j, \ldots)$. In the instance of the two-locus case, every combination $\mathbf{S} = (S_i, S_j)$ is tested for its mutual information $I(\mathbf{S};P)$ with the phenotype, where $\mathbf{S}$ is now a random variable with $3^2 = 9$ possible realizations.[1] Fig. 1 depicts the transition diagram for this case. For each possible zygous state of two SNPs with respective alleles a/A and b/B, the lines correspond to the probabilities of disease onset. From these probabilities, the mutual information between the combination of two SNPs and the phenotype can be calculated.

Analogously, the phenotype can be extended from binary to higher order random variables. This can be helpful for association studies investigating the additional effect of random variables, e.g., for studies categorizing the individuals into cases/controls and smokers/nonsmokers.

The mutual information between genotype and phenotype is a measure for the association of a genetic marker or a combination of markers with the phenotype, both in absolute (through the unit *bits*) and relative terms. However, due to the finite sample size, it is necessary to verify the result in terms of statistical significance. One can show that the mutual information between *independent* random variables when estimated from relative frequencies follows in a very good approximation of a Gamma distribution with parameters $a = (|X|-1)(|Y|-1)/2$ and $b = 1/(N \ln 2)$ (see [14] and Appendix B). In short,

$$I(X;Y) \sim \Gamma\left(\frac{1}{2}(|X|-1)(|Y|-1), \frac{1}{N\ln 2}\right), \tag{8}$$

where N is the sample size and $|X|$ and $|Y|$ denote the numbers of realizations of the random variables X and Y. For example, to determine the significance of $I(S_i;P)$ in a case-control study comprised of N individuals on a significance level α, we check the condition

$$I(S_i;P) \geq \Gamma_{1-\alpha}\left(\frac{1}{2}(2-1)(3-1), \frac{1}{N\ln 2}\right), \tag{9}$$

where $\Gamma_{1-\alpha}(a,b)$ denotes the $(1-\alpha)$-quantile of the Gamma distribution. Since the genotyping failure rate (number of missing values) varies slightly between the individual markers, the critical values need to be calculated for each marker or marker combination separately.

There is a close relationship between mutual information and the standard tests of independence, the χ^2 and log-likelihood ratio tests (see Appendix B). This is not very surprising, considering that the log-likelihood ratio test was proposed in an information theoretic context [15]. Using mutual information, a scaled result in bits is readily obtained, with the advantage that test variables are comparable without prior conversion to their corresponding p-values. The main advantage, however, is that information theory offers a well-defined framework to construct comprehensive gene mapping algorithms, allowing for different types of variables (any type of genetic markers, qualitative and quantitative phenotypes, combined or uncombined markers, and genotype or haplotype probabilities).

3.1 Relevance Chains

In analyzing complex diseases, neither the number and position of susceptibility loci nor their independence in contributing to the disease are known. There may be loci that only modify the effect of main contributing loci. Standard statistical concepts such as stepwise tests of regression models have been applied to unify the procedure for evaluating the additional effects of polymorphisms [16]. Our approach uses the concept of conditional mutual information to determine relevance chains of significantly contributing primary, secondary, and so on variants. Each chain's first element is a single SNP that has significant mutual information with the phenotype. The second and following elements are SNPs that contribute significant *additional* mutual information about the phenotype. Relevance chains are found by the following algorithm:

1. For all $i = 1, \ldots, L$, calculate $I(S_i;P)$. Determine significance by means of comparison with critical value $\Gamma_{1-\alpha}((|S|-1)(|P|-1)/2, 1/(N\ln 2))$.
2. Make each SNP that exhibits significant mutual information the first element of a new chain.
3. In this step, markers are determined that contain *additional* significant information about the phenotype given the markers already found in Step 2. If S_c denotes the SNP that is the first element of relevance chain c, calculate $I(S_i;P|S_c)$ for all $i \neq c$. Compare to critical value $\Gamma_{1-\alpha}(\frac{|S|}{2}(|S|-1)(|P|-1), 1/(N\ln 2))$.
4. If only one additionally significant marker is found, add to the current chain. If k additionally significant markers are found, duplicate the current chain k times and add each additional marker to one of the new chains. Repeat Steps 3 and 4 for each existing relevance chain.
5. Depending on the available sample size, either terminate the algorithm or further extend chains. To do so, start over with Step 3 and let the current chain

1. It is assumed that there is no knowledge about the data's haplotype phase. The proposed method can be easily extended to the case where haplotype effects are suspected and the haplotype phase has been estimated. This is discussed in Section 3.2.

contain m markers. These m markers are combined in a vector random variable $\mathbf{S}$. The mutual information $I(S_i; P|\mathbf{S})$ is calculated for all i and compared to the critical value $\Gamma_{1-\alpha}(\frac{|S|^m}{2}(|S|-1)(|P|-1), 1/(N \ln 2))$.

To understand the evaluation of significance in Step 3, it is important to know that the mutual information between two conditionally independent random variables X, Y given a random variable Z is given by [14]

$$I(X;Y|Z) \sim \Gamma\left(\frac{|Z|}{2}(|X|-1)(|Y|-1), \frac{1}{N \ln 2}\right). \qquad (10)$$

To avoid estimating the probabilities in (6) from too few samples, it is sensible to stop the algorithm after a certain chain length, depending on the sample size N.

The resulting relevance chains are now subject to interpretation. From the biological point of view, an SNP that only appears as a secondary (i.e., not first) element in one or more chains (asymmetrical chains) is unlikely to be directly causal to the disease. However, such a marker does modulate the existing genotype-phenotype relationship of the first SNP. Contrarily, a multilocus disease with equally causal markers is likely to result in symmetrical chains where each causal marker appears as a first element in one chain and as a secondary one in the others. This follows as a consequence from the chain rule of mutual information (7) and the symmetry property (4).

However, there is one constellation where the described algorithm fails to detect the causal markers. Consider the (rather theoretical) example of two causal markers, S_1 and S_2, where $I(S_1; P) = I(S_2; P) = 0$ bit, but $I(S_1, S_2; P) = 1$ bit. As neither marker is found in Step 1 of the algorithm, both markers remain undetected. This example underlines the importance of multiple-loci mutual information as introduced in Section 3, when causal variants are uncorrelated. Such purely epistatic models are described in [17].

3.2 Haplotype Data and Phase-Known Genotypes

The methods described so far investigate the directly observed, unphased genotypes, which, in most cases, is the type of raw information at hand. It is, however, indicated that haplotype analyses may be advantageous in cases where the genotyped loci are not causal themselves, but rather in linkage disequilibrium (LD) with the causal variant or in certain cis-regulated cases where multiple marker alleles contribute to a common disease only when they are located at the same chromosome. In the instance of the two-locus model depicted in Fig. 1, we have 16 different phased haplotype combinations (phase-known genotypes) instead of nine possible genotype combinations. Assuming no parent-of-origin effects (i.e., equal disease risks for genotypes *ij/kl* and *kl/ij*), these 16 states collapse to 10 states (in Fig. 1, the double heterozygote *aAbB* is split into *ab/AB* and *Ab/aB*) [16].

Algorithms have been proposed to infer the haplotype phase statistically from genotype data, if phased data are not directly determined by the experimental design [18], [19], [20]. A discussion of phase estimation on gene mapping from an information theoretic point of view is given in Section 6 and Appendix C.

TABLE 1
Phase-Known Genotype Probabilities as Calculated by the EM Algorithm

	Ab/ab	AB/aB	Ab/aB	AB/ab
control 1	1	0	0	0
control 2	0	0	0.5	0.5
case 1	0	0	0.5	0.5
case 2	0	1	0	0

Generally, the gene mapping of phase-resolved multilocus genotypes is different from the above described procedure insofar as a genotype may have a larger number of realizations, e.g., 10 for two-loci genotypes instead of nine realizations. This increase in the parameter $|\mathbf{S}|$ will change the underlying distribution's shape (parameters $|X|$ in (8) and $|X|$ and $|Z|$ in (10) will increase).

Moreover, the haplotype combinations' probabilities of individuals as output by phase estimation algorithms can easily be included in the calculation of MI. This is achieved simply by calculating the relative frequencies from probability-weighted counts of phased genotypes. As a simplified example, consider a study comprised of $N = 4$ individuals (two cases, two controls) genotyped at two loci. One case and one control individual are heterozyguous in both loci and the other two subjects are heterozyguous at the first locus but homozyguous at the second locus. The standard Expectation-Maximization (EM) algorithm is applied to estimate phase in the double heterozygotes *aAbB* into cis and trans constellation, and the probabilities found are given in Table 1.

The relative frequencies needed to calculate the MI ($p(x, y)$ in (3)) for controls are then given as p(Ab/ab,control) = (1+0)/4 = 0.25, p(AB/aB,control) = 0, p(Ab/aB,control) = 0.125, and p(AB/ab,control) = 0.125. The joint probabilities for cases are calculated similarly.

3.3 Extension to Continuous Phenotypes

Until now, the phenotype has been regarded as a discrete, usually binary, variable. In this section, the concept of gene mapping using mutual information will be extended to quantitative traits.

Obviously, it is possible to quantize the phenotype and afterward treat it as a discrete variable. This strategy will result in a loss of information due to quantization and will hardly supply a sufficient number of samples for each quantized category. It is more appropriate to treat the phenotype as a continuous variable.[2] To derive an expression for the mutual information between a discrete and a continuous random variable, we will, however, in a first step, quantize the phenotype with step size $\Delta > 0$ and obtain its probability mass function (PMF) as

2. Note that, here, P denotes the continuous phenotype variable and p a possible realization, while $p(x)$ and $f(x)$ are PMFs and PDFs, respectively. Hence, $f(p)$ denotes the phenotype's probability density function.

$$p(j\Delta) = \int_{j\Delta}^{(j+1)\Delta} f(p)dp, \quad j = 0, 1, 2, \ldots, \tag{11}$$

and the joint PMF as

$$p(j\Delta, s) = \int_{j\Delta}^{(j+1)\Delta} f(p, s)dp, \quad j = 0, 1, 2, \ldots, \tag{12}$$

where s denote the possible realizations of S, i.e., the SNP's possible allele combinations.

The mutual information between S and P is then

$$I(S; P) = \sum_s \sum_{j=0}^{\infty} p(j\Delta, s) \log_2 \frac{p(j\Delta, s)}{p(j\Delta) \cdot p(s)}. \tag{13}$$

For sufficiently small Δ, the integrals can be approximated by

$$\int_{j\Delta}^{(j+1)\Delta} f(p)dp \approx f(j\Delta) \cdot \Delta \tag{14}$$

and

$$\int_{j\Delta}^{(j+1)\Delta} f(p, s)dp \approx f(j\Delta, s) \cdot \Delta, \tag{15}$$

which turns (13) into

$$I(S; P) = \sum_s \sum_{j=0}^{\infty} f(j\Delta, s) \cdot \Delta \log_2 \frac{f(j\Delta, s) \cdot \Delta}{f(j\Delta) \cdot \Delta \cdot p(s)}. \tag{16}$$

As $\Delta \to 0$, (16) becomes

$$I(S; P) = \sum_s \int_{\mathcal{S}_P} f(p, s) \log_2 \frac{f(p, s)}{f(p) \cdot p(s)} dp, \tag{17}$$

where $\mathcal{S}_P$ denotes the support set of the phenotype variable P. Equation (17) is the mutual information between a discrete and a continuous random variable.

To evaluate (17), two steps are necessary. First, the PDFs $f(p)$ and $f(p, s) \forall s$ must be estimated by means of the available sample data. This is a problem that can hardly be automated and needs to be tackled by inspection. Very often, the PDFs will be normal distributions, in which case, the problem reduces to parameter estimation of the mean μ and variance σ^2. However, other PDFs may be considered as well. In the second step, the integral in (17) has to be evaluated for each part of the sum. In general, this may not be possible analytically. Hence, numerical integration methods need to be applied in order to solve (17) for the mutual information.

In preliminary tests, this method delivered satisfactory results. However, it turns out that particular emphasis must be put on the issue of numerical stability and accuracy in integration.

4 Marker Clustering Using Information Theory

So far, we have used mutual information between phenotype and genotype in a variety of cases. In this section, we use the mutual information between SNPs to find groups or clusters of correlated genetic markers which are likely to form evolutionary entities. This is an important tool for gene mapping as it can give additional hints to which markers are to be jointly interpreted. Multiple study populations with deviating correlation structures (i.e., different marker clusters) may help to pinpoint the potential causal region determined by these correlated marker groups [21].

Our approach is based on rearranging markers in two or three-dimensional space according to their correlation to visualize their interdependence. From these diagrams, the relation between single markers, i.e., their tendency to form clusters, will become obvious. Moreover, the relative positions of neighboring marker groups or clusters may provide helpful knowledge about the long-range LD.

Clusters are formed in three steps. First, the pairwise mutual information between SNPs, $I(S_i; S_j), i, j = 1, \ldots, L$, is determined. In order to avoid biased results, it is sensible to use only population-based controls for this step. Next, the pairwise mutual information, which is a similarity measure, needs to be transformed into a dissimilarity or distance measure. To do so, the *normalized information distance* [22] is used, a metric that satisfies the identity and symmetry axioms and (neglecting a small error term) the triangle inequality. It is defined as

$$\begin{aligned} d(S_i; S_j) &= 1 - \frac{H(S_i) - H(S_i|S_j)}{H(S_i, S_j)} \\ &= 1 - \frac{I(S_i; S_j)}{H(S_i) + H(S_j) + I(S_i; S_j)}. \end{aligned} \tag{18}$$

Calculating all distances, we obtain a diagonally symmetrical $L \times L$ matrix $\mathbf{D}$. In the third step, the method of classical *multidimensional scaling* (MDS) (sometimes called *principal coordinate analysis*) is used to obtain the markers' coordinates in two, three, or n-dimensional space. We used the simple and standard MDS, but other cluster methods could also be employed. The interested reader is referred to [23] for a step-by-step introduction to MDS.

5 Application

5.1 Data Sets

The proposed methods were tested on simulated and real data sets. We used the real data set described in [24]. In this study, 108 SNPs and one microsatellite marker (recoded as biallelic) across a 317 kb region of the genes CD28, CTLA4, and ICOS were genotyped in 384 cases of Graves' autoimmune disease and 652 controls. Genotype-phenotype association has been originally tested by logistic regression analyses mostly assuming a multiplicative model of allelic risks (multiplicative effects of alleles at a locus).

For the case-control simulation, a haplotype population was generated by a coalescent approach allowing for random mutations and recombinations (recombination parameter $4N_e r = 100$ for 100 kb) [25]. Haplotypes were characterized by a sequence of SNP alleles. SNPs with a minor allele frequency of less than 0.05 were excluded and two loci with minor allele frequencies between 0.1 and 0.3 were chosen as the causal variants. A two-locus multiplicative association model with allelic relative risks of 1.5 and a phenocopy rate of 0.01 was specified [26]. This model conferred strong effect amplification among the causal

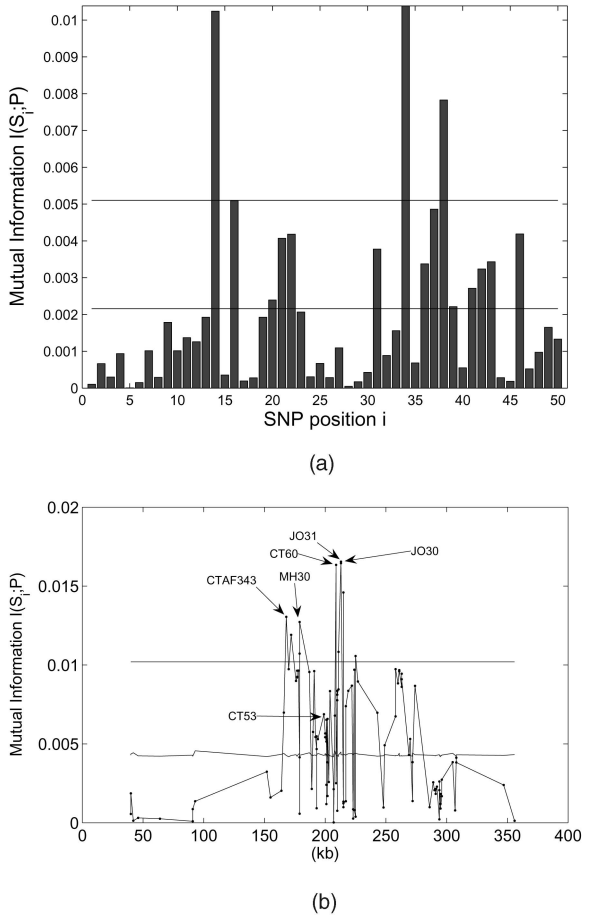

Fig. 2. One-locus mutual information results for the simulated data set (a) and the real autoimmune disease data set (b). The lines indicate the 5 percent significance levels for table-wise testing (upper line) and single-marker testing (lower line).

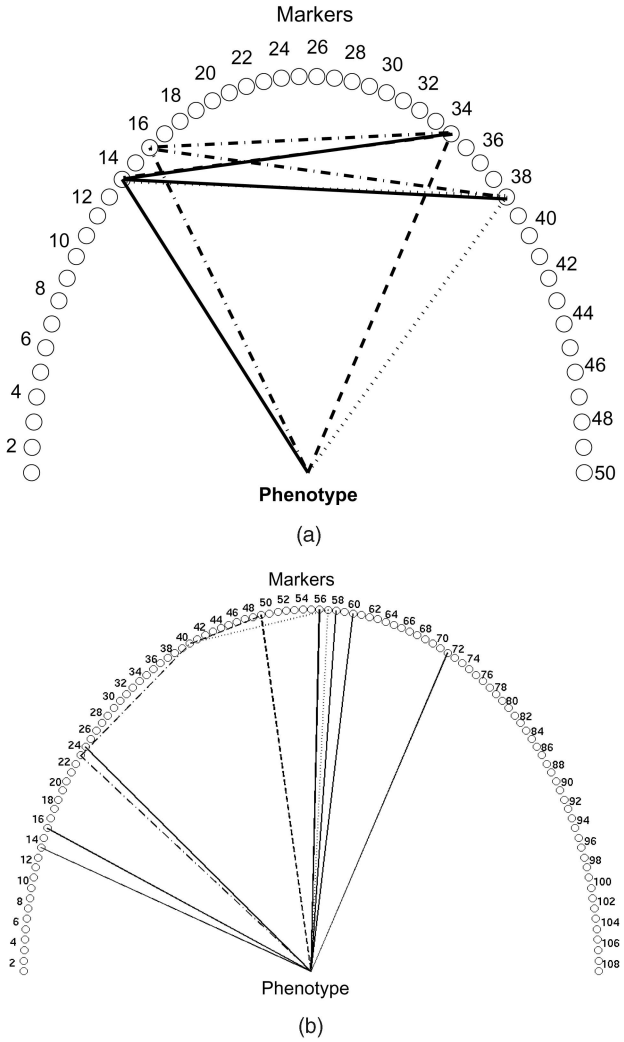

Fig. 3. Relevance chains for simulated data (a) and autoimmune disease data (b).

variants. After removing the causal SNPs (at positions between 11 and 12 and between 37 and 38), a sample of 1,000 case-control pairs was drawn. A genotyping error of rate 0.002 was allowed.

5.2 Results

Fig. 2 shows the results of one-locus mutual information for the simulated and real data sets. It should be noted that the effects measured are relatively weak (≈ 0.01 bit as compared to the theoretical maximum of 1 bit). To determine the results' significance, the critical values (based on $\alpha = 5\%$ significance level) have been determined and plotted. The lower lines represent the analytical-based critical values of single tests (see Section 3), whereas the upper lines represent permutation-based critical values of the total study (global null hypothesis).

Using critical values obtained from permutation tests is indicated when the aim is to fix the overall, i.e., study-wide, type I error probability. This way—although time-consuming—multiple dependent tests can be corrected for with higher accuracy as compared to FDR or Bonferroni corrections. The permutation tests of primary associations are performed by permuting case/control labels. When the significance of relevance chain elements of order 2 or higher (conditional MI) are evaluated, the secondary marker's genotype values need to be permuted in order not to destroy the significant association between the primary markers and the phenotype. It must be noted that, in studies with larger numbers of markers and the application of studywise critical values, the natural trade-off between type I and type II error probabilities will greatly increase the risk for false negative decisions. Therefore, the method for determining the critical values must be chosen carefully. If the emphasis is on minimization or control of false-positive results, it is recommended to use study-wise tests.

To assess the quality of the relevance chain algorithm, the simulated data set with its known properties is used. In this data set, two causal multiplicative markers with approximately equal strength have been simulated at positions 11-12 and 37-38. The analysis of the simulated data set delivers the relevance chains (14, 34), (14, 38), (16, 34), (16, 38), (34, 14), (38, 14) (see Fig. 3).

Such a symmetrical result indicates two equally and jointly causal markers. For example, SNP 34 is a second-order element to SNP 14, but the reverse is also true. Fig. 3 displays the causal regions and their interdependence, verifying that our algorithm indeed delivers correct results

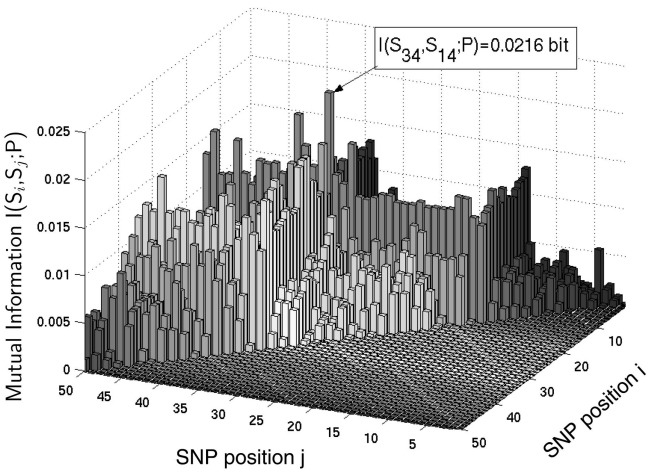

Fig. 4. Two-locus mutual information for simulated data. The maximum value is $I(S_{14}, S_{34}; P) = 0.0216$ bit.

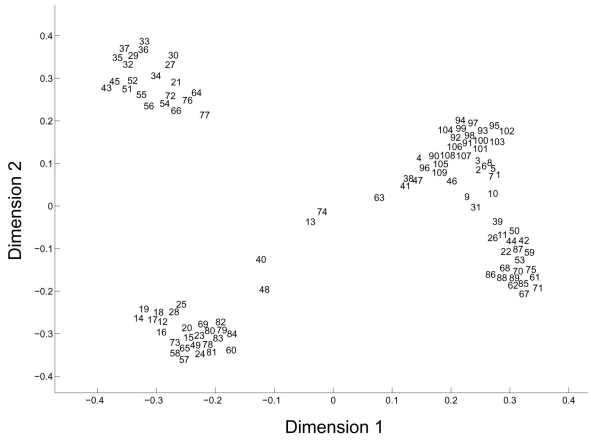

Fig. 5. Marker clustering in a two-dimensional scaling; autoimmune disease data.

in scenarios of multiple causal loci at nonneighboring positions.

In Fig. 4, the mutual information between pairs of markers and the phenotype is depicted for the simulated data set. The maximum value of 0.02 occurs at the combination of SNPs 14 and 34, a result that corresponds to the two maximal one-locus mutual information values shown in Fig. 2. The critical value in this case is $\Gamma_{0.95}((9-1)(2-1)/2, 1/(2000 \cdot \ln 2)) = 0.0056$ bit.

Our analysis of the autoimmune disease data set revealed two study-wise significantly associated regions (see Fig. 2), which are identical to the most promising regions found by the logistic regression analyses reported by [24]. The first region is located 5′ upstream of the gene CTLA4 and comprises the markers 13 (= CTAF343), 15, 23, and 24 (= MH30). The second region at the 3′ end of CTLA4 includes the markers 49 (= CT60), 56, 57 (= JO31), 58 (= JO30), 60 (= JO27_1), and 72. The markers of both regions belong to one cluster with similar patterns of cross-individual variation, except for the weakly associated markers 56 and 72 (see Fig. 5).

This set of markers is likely described by similar evolutionary histories and ages and, most likely, maps only a single causal locus within these two regions. This observation thus corroborates the suggestion of [24] that only one causal variant is located at the 3′ end of CTLA4. In contrast to the original article, however, we found an additional significant signal at SNP marker 40 (= CT53) after adjustment for the main effects at markers 23, 49, and 57 (see Fig. 3). This additional (modifying) association was not detected by a stepwise regression procedure in the original article and may not be attributed to slightly different significance thresholds between our and the original analysis (the p values of the conditional MI between marker 40 and the phenotype given markers 23, 49, and 57 are highly significant with $2 \cdot 10^{-4}$, $2 \cdot 10^{-5}$, and $8 \cdot 10^{-5}$, respectively).

It represents a genuine new result of our method which was missed by other analysis methods. Interestingly, marker 40 lies in the promotor region of CTLA4, which directly corroborates the experimental finding of genetically controlled expression variation in CTLA4 splice forms reported by [24].

6 Discussion and Conclusions

The behavior of genetic associations with complex traits is expected to be influenced by multiple contributing loci, effect modification between loci, and various transmission and interaction systems. The genetic patterns, however, are largely unknown for specific gene-trait association studies. A potential advantage of our approach is its lack of a priori assumptions and the ability to easily extend the method to complex analyses that deal with multiple loci, haplotype probabilities, quantitative traits, and environmental factors. The idea is based on the perception that the association between a genetic marker and a complex disease can be interpreted as a quantity of information contained in the marker about the disease. The concept of relevance chains can then be used to find groups of jointly associated markers and their respective order, corresponding to the conception of genetic effect modifiers. Both a time-saving analytical mode and a permutation-based study-wise mode of significance determination of MI values are implemented in our method.

The results obtained for clinical as well as simulated data sets indicate that our methods constitute a promising approach. The question needs to be discussed of how the proposed methodology relates to commonly used methods. Traditionally, there are two standard statistical methods applied to gene mapping. One method is to test the observed contingency tables for independence by means of the χ^2 or the log-likelihood test (e.g., [27], [4]). These methods are defined very similarly to mutual information (see Appendix B). However, as we have seen in Section 2, the concepts of information theory constitute a background that allows us to go far beyond simple single-marker tests for independence. Moreover, with results scaled to bits, a very intuitive assessment of a marker's causality or strength becomes possible. The other group of methods used for gene mapping are regression-based approaches. Very often, linear or logistic regression is used to analyze the influence of haplotypes, alleles, or genotypes on the phenotype (e.g.,

[3]). A unified stepwise regression procedure very similar to our relevance chain method has been outlined in [16]. It tests the likelihood ratio between a full regression model (e.g., the main effects of a primary and secondary locus) and a constraint model (only the main effects of the primary marker). Using this method for our simulation data, the same modifying (secondary) markers are revealed as in our relevance chains (Fig. 3) when we apply a very conservative significance level ($p < 0.00005$) for each single test. Our method applies a permutation procedure to adjust for testing multiple secondary SNPs and needs no conservative significance level to be defined. As expected, no haplotype-specific effects across primary and secondary markers were found because the two causal loci were simulated as unlinked.

Our approach also revealed very similar results in the real data set example as compared to the original article, except for a new additional study-wise significant signal in the promoter region of CTLA4. In the original article, the effect of secondary loci in addition to the main associated loci was tested, assuming a multiplicative model for the allele effects [24]. Such trend regression approaches, however, imply a continually increasing or decreasing causality scheme across genotypes, which is possibly not always an accurate assumption. The slight difference between the original and our results might be attributed to the fact that the use of mutual information does not assume any particular mode of allelic risk.

Section 3.2 describes how our methods can be easily applied to analyze phase-estimated haplotype data. It may be argued that a haplotype analysis is indicated for certain cis-regulated cases where multiple marker alleles contribute to a common disease only when they are located at the same chromosome. However, in a situation where the number, locations, and functional patterns of these causal markers are largely unknown, this advantage comes at the cost of possibly inaccurate assumptions, e.g., random mating (panmixia) or evolutionary relatedness of haplotypes as assumed by some statistical techniques of haplotype phase estimation. Moreover, the length of analyzed haplotypes has to be determined arbitrarily (sliding window procedures) or by appropriate haplotype block definitions. It is often argued that reconstructing the haplotype phase can increase the statistical power of a test method. From an information theoretic point of view, it is easily shown that this increase in power cannot be regarded as a measure of test quality as even the worst of phase estimation techniques will increase the test power. This matter is discussed in Appendix C.

Applying the simple, yet theoretically well-defined methods from information theory for single or multiple-marker analysis and the determination of relevance chains seems to deliver satisfactory results, while assumptions such as HWE, random mating, haplotype length and phase, linear association models, or number and position of causal markers do not have to be made. We think, therefore, that the representation of marker-marker and marker-phenotype relationships with one simple and basic measure of information lays out a consistent framework for a first screen in gene mapping approaches.

Appendix A

Notation

For reasons of reading clarity, the notation used in this paper is rather compact. Capital letters such as S denote random variables, while small letters such as s denote an outcome of a random experiment and, hence, a realization of S. These realizations s are elements of the support set $\mathcal{S}_S$, whose size is denoted by $|S|$. The notation $\sum_s$ means a sum over all realizations of S, correctly $\sum_{s\in\mathcal{S}_S}$. Probability mass functions (PMF) are denoted by $p(\cdot)$ and probability density functions (PDF) are denoted by $f(\cdot)$. Consequently, $p(s)$ means $p_S(s)$, $p(p,s)$ means $p_{PS}(p,s)$, etc.

$\Gamma(a,b)$ is the Gamma distribution with shape parameter a and scale parameter b. $\Gamma_{1-\alpha}(a,b)$ denotes the $(1-\alpha)$-quantile of the Gamma distribution, i.e., that value that is exceeded with probability α [28].

Appendix B

The Connection between χ^2, Log-Likelihood Ratio, and Mutual Information

Without going into too much detail, the connection between mutual information and the χ^2 and log-likelihood ratio (or G^2 or $2\hat{I}$) tests is discussed. Using the definition of relative frequencies,

$$p(\text{event}) = \frac{\text{event counts}}{\text{total counts}}, \tag{19}$$

it can be easily shown that

$$2\hat{I} = 2N \ln 2 \cdot I(X;Y), \tag{20}$$

where N is the number of observed samples (the total counts).

The χ^2 and log-likelihood ratio tests are very similar; in fact, the χ^2 test is a second-order Taylor approximation of $2\hat{I}$. For independent random variables, both have an asymptotic χ^2 distribution. (Very often in the literature, $2\hat{I}$ is said to be asymptotically distributed as χ^2. This, however, is only approximately correct, viz. for its approximation, the χ^2 test.)

As mentioned in the text, the logarithm base b used in entropy expressions is arbitrary and the result is equal bar a scale factor $\ln b$. Using the natural logarithm, we can expand the expression for mutual information $I(X;Y)$ into a Taylor series about expansion point $p_{XY} \equiv p_X \cdot p_Y$, i.e., "about independence," and obtain

$$I(X;Y) \approx \frac{1}{2}\sum_x \sum_y \frac{(p(x,y) - p(x)p(y))^2}{p(x)p(y)}. \tag{21}$$

Obviously, this expression relates to the χ^2 test with the same constant factor $2N$ (and an additional factor $\ln 2$ if the binary logarithm is used). The direct proof that (21) has a Gamma distribution is rather involved. However, the same fact can be quite easily derived from knowing the χ^2 test variable follows a χ^2 distribution (given the null hypothesis is true). Since $I = \frac{X^2}{2N\ln 2}$, we can scale the χ^2 distribution by the factor $2N \ln 2$ and obtain a Gamma distribution.

One particularly important consequence of the relation between χ^2 and MI is that all accuracy properties of the χ^2 test can be attributed to MI, too. Most important for the proposed gene mapping methods is the fact that the χ^2 distribution (and, thus, our gamma distribution, too) is considered a very good approximation for sample sizes as small as $N = 10$.

For a more detailed discussion of the matter, the reader is referred to [14].

APPENDIX C
HAPLOTYPE ESTIMATION AND TEST POWER

The methods proposed in this paper can be easily used for the analysis of data whose haplotype phase has been reconstructed by statistical methods such as the EM algorithm [18]. Haplotype phase estimation is often recommended with the argument that it can increase the statistical test power defined as $1 - \beta$, where β is the type II (false negative) error probability.

To understand the effect of haplotype phase estimation from an information theoretic point of view, we consider the reverse process. Let a random variable G denote an SNP sequence (length $l \geq 1$) with unresolved gametic phase and let H denote the same sequence with estimated phase. Obviously, G is easily derived from H by combining several heterozygous states into one. For instance, the genotypic state $aAbB$ in Fig. 1 would split into the haplotype combinations ab/AB, aB/Ab, Ab/aB, and AB/ab.

Hence, G is a function of H. This implies that G and any random variable P (e.g., the phenotype) are conditionally independent given H. In statistical terms, P, H, and G form a Markov chain in that order. Under this condition, one of the fundamental theorems of information theory, the *data processing inequality* [6], states that

$$I(G;P) \leq I(H;P). \tag{22}$$

Thinking in the other direction, this proves that even the worst of haplotype phase estimation techniques, e.g., guessing, will increase the mutual information between the SNP sequence and the phenotype. Depending on the quality of the phase estimation algorithm, this increase in mutual information will be more or less justified. However, since the haplotype phase can never be reconstructed unambiguously, a certain amount of mutual information will always be unjustified or spurious information.

The increase in mutual information will always increase the test power, $1 - \beta$, since the increase in mutual information makes it more likely for a marker set to lie above the significance threshold. Hence, the type II or false negative error probability β decreases and the test power increases. However, this test power increase is obtained at the cost of more type I errors because the spurious information will make it more likely to assess a noncausal SNP sequence as causal, hence making a false positive decision.

This shows that an increase in test power cannot be taken as a quality measure for haplotype phase estimation techniques. When phase estimated data is analyzed, it needs to be kept in mind that part of the observed associations may be spurious ones.

ACKNOWLEDGMENTS

Jakob C. Mueller and Christophe Andreoli were supported by the BFAM (Bioinformatics for the functional analysis of mammalian genomes) and NGFN (National Genome Research Network) projects from the German Federal Ministry of Education and Research. The authors thank John A. Todd, Neil Walker, and the JDRF/WT DIL for supplying the raw data set analyzed in [24].

REFERENCES

[1] D. Botstein and N. Risch, "Discovering Genotype Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease," *Nature Genetics,* vol. 33 (suppl.), pp. 228-237, Mar. 2003.

[2] L. Cardon and J. Bell, "Association Study Designs for Complex Diseases," *Nature Rev. Genetics,* vol. 2, no. 2, pp. 91-99, Feb. 2001.

[3] D. Zaykin, P. Westfall, S. Young, M. Karnoub, M. Wagner, and M. Ehm, "Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals," *Human Heredity,* vol. 53, no. 2, pp. 79-91, May 2002.

[4] D. Clayton, "Population Association," *Handbook of Statistical Genetics,* D. Balding, M. Bishop, and C. Cannings, eds., pp. 519-540, Chichester: John Wiley & Sons, 2001.

[5] C.E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical J.,* vol. 27, pp. 379-423, 623-656, July-Oct. 1948.

[6] T. Cover and J. Thomas, *Elements of Information Theory.* New York: John Wiley & Sons, 1991.

[7] A. Butte and I. Kohane, "Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements," *Proc. Pacific Symp. Biocomputing,* pp. 418-429, Jan. 2000.

[8] J. Kasturi, R. Acharya, and M. Ramanathan, "An Information Theoretic Approach for Analyzing Temporal Patterns of Gene Expression," *Bioinformatics,* vol. 19, no. 4, pp. 449-458, 2003.

[9] M. Nothnagel, R. Fürst, and K. Rohde, "Entropy as a Measure for Linkage Disequilibrium over Multilocus Haplotype Blocks," *Human Heredity,* vol. 54, pp. 186-198, 2003.

[10] J. Hampe, S. Schreiber, and M. Krawczak, "Entropy-Based SNP Selection for Genetic Association Studies," *Human Genetics,* vol. 114, no. 1, pp. 36-43, Dec. 2003.

[11] I. Grosse, H. Herzel, S. Buldyrev, and H. Stanley, "Species Independence of Mutual Information in Coding and Noncoding DNA," *Physical Rev. E,* vol. 61, no. 5, pp. 5624-5629, 2000.

[12] A. Tsalenko, A. Ben-Dor, N. Cox, and Z. Yakhini, "Methods for Analysis and Visualization of SNP Genotype Data for Complex Diseases," *Proc. Pacific Symp. Biocomputing,* vol. 8, pp. 548-561, Jan. 2003.

[13] J. Mueller, E. Bresch, Z. Dawy, T. Bettecken, T. Meitinger, and J. Hagenauer, "Shannon's Mutual Information Applied to Population-Based Gene Mapping," *Am. J. Human Genetics,* vol. 73, no. 5 (suppl.), p. 610, Nov. 2003.

[14] B. Goebel, Z. Dawy, J. Hagenauer, and J. Mueller, "An Approximation to the Distribution of Finite Sample Size Mutual Information Estimates," *Proc. IEEE Int'l Conf. Comm.,* May 2005.

[15] S. Kullback, *Information Theory and Statistics.* New York: John Wiley & Sons, 1959.

[16] H. Cordell and D. Clayton, "A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data: Application to HLA in Type 1 Diabetes," *Am. J. Human Genetics,* vol. 70, no. 1, pp. 124-141, Jan. 2002.

[17] R. Culverhouse, B. Suarez, J. Lin, and T. Reich, "A Perspective on Epistasis: Limits of Models Displaying No Main Effect," *Am. J. Human Genetics,* vol. 70, no. 2, pp. 461-471, Feb. 2002.

[18] L. Excoffier and M. Slatkin, "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population," *Molecular Biology and Evolution,* vol. 12, no. 5, pp. 921-927, Sept. 1995.

[19] M. Stephens, N. Smith, and P. Donnelly, "A New Statistical Method for Haplotype Reconstruction from Population Data," *Am. J. Human Genetics,* vol. 68, no. 4, pp. 978-989, Apr. 2001.
[20] D. Fallin and N. Schork, "Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data," *Am. J. Human Genetics,* vol. 67, no. 4, pp. 947-959, Oct. 2000.
[21] J. Mueller, E. Lohmussaar, R. Magi, M. Remm, T. Bettecken, P. Lichtner, S. Biskup, T. Illig, A. Pfeufer, J. Luedemann, S. Schreiber, P. Pramstaller, I. Pichler, G. Romeo, A. Gaddi, A. Testa, H. Wichmann, A. Metspalu, and T. Meitinger, "Linkage Disequilibrium Patterns and TagSNP Transferability among European Populations," *Am. J. Human Genetics,* vol. 76, no. 3, pp. 387-398, Mar. 2005.
[22] M. Li, X. Chen, X. Li, B. Ma, and P. Vityi, "The Similarity Metric," *Proc. 14th Ann. ACM-SIAM Symp. Discrete Algorithms,* pp. 863-872, 2003.
[23] T. Cox and M. Cox, *Multidimensional Scaling.* London: Chapman & Hall, 1994.
[24] H. Ueda, J. Howson, L. Esposito, J. Heward, H. Snook, G. Chamberlain, D. Rainbow, K. Hunter, A. Smith, G.D. Genova, M. Herr, I. Dahlmand, F. Payne, D. Smyth, C. Lowe, R. Twells, S. Howlett, B. Healy, S. Nutland, H. Rance, V. Everett, L. Smink, A. Lam, H. Cordell, N. Walker, C. Bordin, J. Hulme, C. Motzo, F. Cucca, J. Hess, M. Metzker, J. Rogers, S. Gregory, A. Allahabadia, R. Nithiyananthan, E. Tuomilehto-Wolf, J. Tuomilehto, P. Bingley, K. Gillespie, D. Undlien, K. Ronningen, C. Guja, C. Ionescu-Tirgoviste, D. Savage, A. Maxwell, D. Carson, C. Patterson, J. Franklyn, D. Clayton, L. Peterson, L. Wicker, J. Todd, and S. Gough, "Association of the T-Cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease," *Nature,* vol. 423, no. 6939, pp. 506-511, May 2003.
[25] R. Hudson, "Generating Samples under a Wright-Fisher Neutral Model of Genetic Variation," *Bioinformatics,* vol. 18, pp. 337-338, Feb. 2002.
[26] M. Nothnagel, "Simulation of LD Block-Structured SNP Haplotype Data and Its Use for the Analysis of Case-Control Data by Supervised Learning Methods," *Am. J. Human Genetics,* vol. 71 (suppl.), no. A2363, Oct. 2002.
[27] B. Everitt, *The Analysis of Contingency Tables.* London: Chapman and Hall, 1977.
[28] H. Sahai and A. Khurshid, *Pocket Dictionary of Statistics.* McGraw-Hill/Irwin, 2002, http://www.mhhe.com/business/opsci/bstat/keyterm.mhtml.

Zaher Dawy received the BE degree in computer and communications engineering with high distinction from the American University of Beirut (AUB) in 1998. He received the MSc and Dr.-Ing. degrees in electrical engineering with excellent distinction from Munich University of Technology (TUM) in 2000 and 2004, respectively. Between 1999 and 2000, he worked as a part-time communications engineer at Siemens AG research labs in Munich, focusing on the development of enhancement techniques for UMTS. At TUM, between 2000 and 2003, he managed a research project with Siemens AG where he designed advanced multiuser receiver structures for UMTS base stations. Since September 2004, he has been an assistant professor in the Electrical and Computer Engineering Department at the American University of Beirut. His research interests are in the areas of wireless communications, multiuser information theory, and computational biology. He is a member of the IEEE.

Bernhard Goebel received the Diplom-Ingenieur degree in electrical engineering from Munich University of Technology (TUM) in 2004 with a master's thesis on applications of information theory to gene mapping. In 2002-2003, he spent graduate semesters at the University of Southampton, United Kingdom, and Siemens Corporate Research, Princeton, New Jersey, doing a research internship in the field of medical imaging. Currently, he is a research assistant and a PhD student in optical communications at TUM working on information-theoretic optimization of fiber-optic communication systems. He is a student member of the IEEE.

Joachim Hagenauer received his degrees from the Technical University of Darmstadt, Germany, where he served as an assistant professor. He held a postdoctoral fellowship position at the IBM T.J. Watson Research Center, Yorktown Heights, New York, a one-year visiting position year at AT&T Bell Laboratories, Crawford Hill, and a research position at the German Aerospace Center (DLR), Oberpfaffenhofen. Since 1990, he has been the director of the DLR Institute for Communications Technology. Since April 1993, he has been a full professor of telecommunications at the Munich University of Technology (TUM) and, since 2002, he has been a full member of the Bavarian Academy of Science. Professor Hagenauer is a fellow of the IEEE, the recipient of the 1996 E.H. Armstrong Award of the IEEE Communications Society and of the IEEE 2003 Alexander Graham Bell Medal. In 2001, he served as the president of the IEEE Information Theory Society.

Christophe Andreoli Photograph and biography not available at time of publication.

Thomas Meitinger qualified in medicine and biology at the Ludwig-Maximilians-University Munich. He was a visiting scientist at the Genetics Laboratory, University of Oxford, between 1985 and 1988. In 1989, he was assigned the position of senior scientist in the Department of Medical Genetics, Ludwig-Maximilians-University of Munich. In 2000, he was appointed the director of the newly founded Institutes of Human Genetics at both the Technical University of Munich and GSF National Research Centre for Environment and Health. His main research interests are focused on gene mapping and functional analyses of disease associated genes. He combines these research activities with the study of mitochondrial genes, proteins, and diseases. A database of mitochondrial proteins (MITOP2) has been set up at the Institute. More recently, he developed an interest in population genetics and is contributing to genetic epidemiological studies carried out at the GSF.

Jakob C. Mueller received the PhD degree from the Department of Population Biology at the Johannes Gutenberg-University in Mainz, Germany, where he served as an assistant professor. Since 2002, he has held a position as a senior scientist at the Institute of Human Genetics (GSF, Munich), at the Institute for Medical Statistics and Epidemiology, and at the Institute for Psychiatry and Psychotherapy at the Technical University in Munich. He is also affiliated with the National Center for Genetic Epidemiological Methods and the Hertie-Institute for Clinical Brain Research, Tuebingen. His research interests include population and evolutionary genetics in animals and humans as well as gene mapping for complex traits through linkage and association studies.